TOPOLOGICAL PRESSURE AND CODING SEQUENCE DENSITY ESTIMATION IN THE HUMAN GENOME

DAVID KOSLICKI AND DANIEL J. THOMPSON

Abstract. Inspired by concepts from ergodic theory, we give new insight into coding sequence (CDS) density estimation for the human genome. Our approach is based on the introduction and study of topological pressure: a numerical quantity assigned to any finite sequence based on an appropriate notion of 'weighted information content'. For human DNA sequences, each codon is assigned a suitable weight, and using a window size of approximately 60,000bp, we obtain a very strong positive correlation between CDS density and topological pressure. The weights are selected by an optimization procedure, and can be interpreted as quantitative data on the relative importance of different codons for the density estimation of coding sequences. This gives new insight into codon usage bias which is an important subject where long standing questions remain open. Inspired again by ergodic theory, we use the weightings on the codons to define a probability measure on finite sequences. We demonstrate that this measure is effective in distinguishing between coding and non-coding human DNA sequences of lengths approximately 5,000bp. We emphasize that topological pressure is a flexible tool and we expect it to be useful for the investigation of many other features of DNA sequences such as interspecies comparison of codon usage bias. We give a first result in this direction, investigating CDS density in the mouse genome and comparing our results with those for the human genome.

1. Introduction

The theory of symbolic dynamical systems is a rich and only partially explored source of techniques for genomic analysis. We introduce a new tool called *topological pressure* (or simply *pressure*), which we apply to the study of the human genome. Pressure can be interpreted as a weighted measure of complexity and is the natural generalization of topological entropy for finite sequences introduced by the first named author in [24]. The primary goals of our analysis are to demonstrate how pressure can predict the distribution of coding sequences across a genome and to use this to recover quantitative data on codon usage bias. This could shed light on the issue of mammalian codon bias, where it is recognized that a complete understanding has not yet been achieved [7, 20, 33].

Date: September 28, 2011.
D.K. is supported by NSF grant DMS-1008538.
D.T. is supported by NSF grant DMS-1101576.

Topological pressure is a well known and well studied concept in the ergodic theory of dynamical systems. The standard version is a quantity associated to a topological dynamical system which measures weighted orbit complexity of the system [31, 32, 42]. We introduce a finite implementation of topological pressure which can be interpreted as a measure of weighted information content of a finite sequence. The pressure of a finite sequence is given by counting (with weights) all subwords of an exponentially shorter length that appear in the original word. Each subword is weighted through the use of a function φ, which we call the *potential*. We focus on potentials which depend on only 3 symbols, so φ is essentially a choice of weighting for each codon. Pressure detects a trade-off between complexity in the sequence and frequency of occurrence of 'favored' codons. This intuition is made rigorous by the Variational Principle from ergodic theory [42], which we recall in §5.2. If a potential can be found so that the pressure correlates strongly with an observed biological phenomena, this gives evidence for the biological importance of those codons which are weighted strongly by that potential.

Our main focus is the selection of potentials for which the topological pressure correlates strongly with the observed distribution of coding sequences when using windows of size approximately $60,000$bp. We optimize this correlation independently on each chromosome with respect to the parameters of the potential and find that the Pearson's correlation coefficient between the coding sequence density and the topological pressure is above 0.9. The parameters obtained on each chromosome are close together (at least for the autosomes) so we average them to obtain a 'canonical' potential for CDS density estimation which we denote by φ_{hs}. We check that a similar potential is obtained when we optimize the correlation across the whole human genome simultaneously. We give a detailed analysis of the pattern of codon weights given by the potential φ_{hs} and observe a number of striking qualitative features that contribute to the investigation of codon usage bias.

Recent research [33, 7, 6, 8, 12, 30, 38] has focused on analyzing the nuanced and oft-debated question of the nature and cause of codon usage bias. While many studies have successfully analyzed a particular influence on synonymous codon usage (context dependency [13], GC content [8], tRNA adaptation [12], etc.), a comprehensive understanding of codon bias (particularly in mammals) remains a challenging open problem. The difficulty in analyzing codon bias can be attributed partly to an over-abundance of plausible statistical and theoretical approaches which are often mutually contradictory [33]. Furthermore, there is general disagreement on how to properly take into consideration features of the sequence such as GC content and context dependencies, as well as the inherent randomness of nucleotide composition. For example, it has been argued that the codon adaptation index [33] should [36] and should not [19] be used to determine the influence of synonymous codon usage on gene expression levels.

The advantage of utilizing topological pressure is its relative simplicity. The definition is entirely combinatorial and implicitly takes account of important considerations such as neighboring dependencies, different choices of reading frame, autocorrelation, background codon frequencies, and GC content. Furthermore, for sequences of suitably large scale, the definition is robust enough to absorb the inherent randomness and noise in nucleotide composition.

These advantages allow us to compare synonymous codon usage between species and its relationship with CDS density estimation. Using the potential φ_{hs}, we show that pressure has good correlation with the CDS distribution of *mus musculus*. In addition, we optimize the correlation of CDS density with pressure over the mouse genome. We observe that the potential obtained this way shares many qualitative features with φ_{hs}. This gives evidence that the parameters in φ_{hs} are biologically meaningful and is a first step in the investigation of interspecies codon usage via topological pressure.

Inspired once more by the techniques of ergodic theory, we demonstrate that any potential φ canonically defines a probability measure on finite sequences via the Variational Principle. This measure, called the *equilibrium measure for* φ, reflects the properties of the potential and can be used to analyze sequences that are orders of magnitude shorter than those on which pressure is utilized. This represents a strategy in which large scale information (pressure) can be utilized to extract information at a much smaller scale (measure of a sequence). The development of robust techniques that detect the coding potential of short sequences is an important area of research [10, 14, 16, 18, 26, 27, 37, 43] with applications to sequence annotation as well as gene prediction. It has been recognized that measures of coding potential based on single sequence nucleotide composition [27, p.i281] are an important part of the problem of differentiating between short reads of coding and non-coding sequences and are complementary to the very effective comparative techniques developed in, for example, [43]. We contribute to this line of research by showing that the equilibrium measure associated with φ_{hs} can effectively distinguish between randomly selected introns and exons in the human genome.

The layout of the paper is as follows: In §2, we define topological pressure for finite sequences. In §3, we investigate the correlation of topological pressure and CDS density in the human genome. In §4, we briefly investigate applications of topological pressure to the mouse genome. In §5, we demonstrate how topological pressure defines a measure on finite sequences, and show that this measure can distinguish between coding sequences and non-coding sequences.

2. Topological pressure for finite sequences

We rigorously develop our implementation of topological pressure for any finite sequence. Let $\mathcal{A}$ be our alphabet, that is, a finite collection of symbols.

Since our application is to the study of DNA sequences, we mainly consider the alphabet $\mathcal{A} = \{A, C, T, G\}$. We consider various spaces of sequences on the alphabet $\mathcal{A}$. We denote the space of sequences of length n by $\mathcal{A}^n$, the space of finite sequences (of any length) $\mathcal{A}^{<\mathbb{N}}$, the space of finite sequences of length at least n by $\mathcal{A}^{\geq n}$ and the space of infinite sequences by $\Sigma = \mathcal{A}^{\mathbb{N}}$. For $w = (w_1, w_2, \dots) \in \Sigma$ or $w = (w_1, w_2, \dots, w_m) \in \mathcal{A}^{\geq n}$, let w_1^n denote the finite word $(w_1, \dots, w_n)$. For $w \in \mathcal{A}^n$, let $[w]$ be the set of sequences $v \in \Sigma$ so that $v_1^n = w$. Let σ be the shift map: For $w = (w_1, w_2, w_3, \dots) \in \mathcal{A}^{\geq 2} \cup \Sigma$, $\sigma((w_1, w_2, w_3, \dots)) = (w_2, w_3, \dots)$.

When we consider norms of matrices $M = (m_{ij})$ and vectors $v = (v_i)$, we consistently use the sum norm, so that $\|M\| = \sum_{i,j} |m_{ij}|$ and $\|v\| = \sum_i |v_i|$.

Definition 2.1. *We say a function ψ on Σ (or on $\mathcal{A}^{\geq m}$) depends on the first m symbols of a word if*

1. *For all $v \in \mathcal{A}^m$, the restriction of ψ to $[v]$ is a constant function.*
2. *There exists $w \in \mathcal{A}^{m-1}$ for which the restriction of ψ to $[w]$ is not a constant function.*

If ψ depends on m symbols, then for $v \in \mathcal{A}^m$, we write $\psi(v)$ for the common value of ψ on $[v]$.

We define topological pressure for finite sequences $w \in \mathcal{A}^{\leq \mathbb{N}}$. Define

$$SW_n(w) = \{u : |u| = n \text{ and } u \subset w\}.$$

The definition depends on the cardinality of the alphabet. To keep the presentation close to our applications, we give the definitions under the assumption that $\#\mathcal{A} = 4$. For alphabets of different cardinality, we simply replace the occurrences of 4 with $\#\mathcal{A}$.

Definition 2.2. *For a word w such that $|w| = 4^n + n - 1$ and a potential function ψ which depends on m symbols, where $n \geq m$, we define the topological pressure of ψ on w to be*

$$P(w, \psi) = \frac{1}{n} \log_4 p(w, \psi), \tag{2.1}$$

where

$$p(w, \psi) = \sum_{u \in SW_n(w)} \exp \sum_{i=0}^{n-m} \psi(\sigma^i u). \tag{2.2}$$

We denote the greatest topological pressure for such words by

$$P_{\max}(n, \psi) = \max\{P(w, \psi) : |w| = 4^n + n - 1\}. \tag{2.3}$$

Remark. When $\psi = \log \varphi$ (log denotes natural logarithm) for a function $\varphi > 0$,

$$P(w, \log \varphi) = \frac{1}{n} \log_4 \left(\sum_{u \in SW_n(w)} \prod_{i=0}^{n-m} \varphi(\sigma^i u) \right). \tag{2.4}$$

We extend this definition to words of an arbitrary finite length.

Definition 2.3. *For a word w with $4^n + n - 1 \leq |w| < 4^{n+1} + n$, we define the topological pressure of ψ on w to be*

$$P(w, \psi) = P(w_1^{4^n+n-1}, \psi). \tag{2.5}$$

That is, the pressure of ψ on w is defined to be the pressure of ψ on the first $4^n + n - 1$ symbols of w.

Remark. An elementary argument given in [24] shows that for each n, there exists a word v^n of length $4^n + n - 1$ which has every word of length n as a subword. It follows that $P_{\max}(\psi, n) = P(v^n, \psi)$ for any function ψ.

Remark. When $\psi = 0$, (2.1) reduces to the definition of topological entropy for finite sequences due to the first named author in [24]. The reason we take the logarithm in base 4 in (2.1) is so that $P_{\max}(n, 0) = 1$.

2.1. **Normalization of potentials.** An arbitrary potential $\psi = \log \varphi$ can be normalized by the addition of a constant. This is useful for a number of reasons, and does not affect the quantities associated to pressure that we study in this paper, such as the equilibrium measures introduced in §5 and correlation with the CDS density developed in §3. For any $t > 0$, we have the formula

$$P(w, \log t\varphi) = \frac{n - m}{n} \log_4 t + P(w, \log \varphi). \tag{2.6}$$

This allows for a variety of normalizations. For us, the most useful normalization is to let $t = \|\varphi\|^{-1}$ so that $t\varphi$ is described by a probability vector. We use this normalization frequently in §2.3.

2.2. **Interpretation of high pressure sequences.** A sequence with high pressure has a good mix of complexity and frequency of 'favored' codons. When using the 0 potential (i.e. entropy), we simply detect high complexity. In [24], it was shown that an intron region of a DNA sequence tends to have higher entropy than an exon region. This is due to the exons having more structure (and hence less randomness). However, for windows of larger size, which may contain numerous intron and exon regions, entropy is a poor indicator of CDS density (see figure 1). In §3.5, we demonstrate that with an appropriate choice of potential, the high pressure sequences correlate very well with those with high coding sequence density. Further insight into the meaning of pressure is given by the Variational Principle from ergodic theory, which we recall in §5.2. The Variational Principle makes precise the intuition that high pressure sequences are those that balance high complexity against high frequency of favored codons.

2.3. **Selection of the Potential.** Two perspectives can be taken regarding selection of the potential φ. The first perspective is to obtain a potential via maximizing the correlation of topological pressure with a given set of biological data. We take this approach in section §3 to select potentials

based on the correlation of pressure with the probability distribution of known coding sequences.

The second perspective is to construct a potential based on known biological phenomena and then utilize topological pressure to analyze the desired feature. Next, we give an example of such a potential.

2.4. **A 1-parameter family of examples.** We give a simple family of examples to illustrate the role of pressure. We write down a potential adapted to detecting regions with high GC content. Since we focus on a much broader and more sophisticated class of potentials in the rest of this paper, this example should be understood as an illustrative toy model. Let $\mathbf{1}_A$ denote the characteristic function of $[A]$, and suppose that $|w| = 4^n + n - 1$. Consider the family of functions

$$\varphi_t = \mathbf{1}_A + \mathbf{1}_T + t(\mathbf{1}_G + \mathbf{1}_C),$$

where $t > 0$. Then

$$p(w, \log \varphi_t) = \sum_{u \in SW_n(w)} \prod_{i=0}^{n-1} \varphi_t(\sigma^i u) = \sum_{u \in SW_n(w)} t^{GC(u)},$$

where $GC(u)$ denotes the total number of occurrences of G and C in the word u. Then

$$P(w, \log \varphi_1) = H_{top}(w)$$

and as t increases from 1, $P(w, \log \varphi_t)$ gives a measure of complexity which assigns increasing importance to sequences with greater GC content. In §3.10, we investigate this family of potentials and how best to choose t in the context of CDS density estimation in the the human genome.

3. Topological Pressure and CDS density estimation

We show that pressure can be used as an effective predictor of coding sequence density for the human genome. The challenge is to make a good choice of potential function. The discovery of a potential function which correlates well with coding sequence density then yields biologically relevant information on the roles of different codons.

3.1. **Coding Sequence Density of the Human Genome.** The *coding sequence density* is the probability density function representing the percentage of coding sequences versus non-coding sequences in non-overlapping windows of a given size. We introduce some notation in order to define the coding sequence density precisely.

Notation 3.1. *Let* $\mathrm{Chr}(i)$ *denote the string which represents the* i^{th} *chromosome of the human genome, and* $\mathrm{Chr}(i, [n, m])$ *denote the substring which starts at position* n *and ends at position* m*. For convenience, we refer to the X and Y chromosomes as the* 23^{rd} *and* 24^{th} *chromosomes respectively.*

We utilize the NCBI hg18 build 36.3 with coding sequences defined by NCBI RefSeq genes and accessed via Wolfram's Mathematica 8.0 [44]. We choose a chromosome and fix an integer window size m to divide the chromosome into non-overlapping windows of length m. The most suitable window sizes for comparision with topological pressure are those of the form $m = 4^n + n - 1$.

Definition 3.1. *For some fixed window size $m = 4^n + n - 1$, we define*

$$\#CS(i, n, x) := \#\{\text{Known coding sequences with initial nucleotide contained in } \mathrm{Chr}(i, [xm + 1, xm + m])\},$$

assuming the chromosome is read in the p to q direction. The coding sequence density is defined to be

$$\mathrm{CDS}(i, n, x) := \#CS(i, n, x)/\#CS(i),$$

where $\#CS(i) := \#\{\text{Known coding sequences in } \mathrm{Chr}(i)\}$.

Thus, the indices i and n tell us to look at the i^{th} chromosome using a window of size $4^n + n - 1$, and the index x describes the starting point of the window along the given chromosome. For fixed i and n, $\mathrm{CDS}(i, n, x)$ is a probability density function of x.

3.2. **Topological Pressure of the Human Genome.** We now set up notation for our application of topological pressure to the human genome.

Definition 3.2. *For a potential $\varphi > 0$ and $m = 4^n + n - 1$, let*

$$P^{\mathrm{hs}}(i, n, x, \varphi) := P(\mathrm{Chr}(i, [mx + 1, mx + m]), \log \varphi),$$

where $P(\cdot, \cdot)$ is the topological pressure defined in equation (2.5).

Thus, $P^{\mathrm{hs}}(i, n, x, \varphi)$ is the topological pressure associated to the x^{th} window of size $4^n + n - 1$ on the human chromosome i, using the potential $\log \varphi$.

3.3. **Selection of φ via maximum correlation with CDS density.** For fixed i and n, we consider $\mathrm{CDS}(i, n, x)$ and $P^{\mathrm{hs}}(i, n, x, \varphi)$ as functions in x. We use the Nelder-Mead [29] method to maximize the correlation between $P^{\mathrm{hs}}(i, n, x, \varphi)$ and $\mathrm{CDS}(i, n, x)$ with respect to potentials φ which depend on 3 symbols. Due to (2.6), we can without loss of generality restrict our attention to the set of potentials whose parameters sum to 1. We thus obtain a set of potentials $\varphi^i_{\max}$ whose associated pressure correlates very well with the coding sequence distribution in the chromosome $\mathrm{Chr}(i)$. We expand on our methodology below, and then present and analyze our results.

3.4. **Methodology.** Considered as functions in x, both $\mathrm{CDS}(i,n,x)$ and $P^{\mathrm{hs}}(i,n,x,\varphi)$ are inherently noisy due to random fluctuations in nucleotide composition in a given chromosome as well as due to incomplete knowledge regarding coding sequences (eg. incorrectly annotated sequences). The noise in both functions is easily suppressed by utilizing a Gaussian filter (convolution with a Gaussian kernel of radius r). We checked that other standard smoothing techniques lead to similar results, and chose the Gaussian filter for our analysis due to its simplicity and speed of implementation. The filter is applied after removing from both $\mathrm{CDS}(i,n,x)$ and $P^{\mathrm{hs}}(i,n,x)$ those x where $\mathrm{Chr}(i,[xm+1,xm+m])$ contained any symbols besides $\{A,C,T,G\}$. The radius of the Gaussian filter is chosen so that $\mathrm{CDS}(i,n,x)$ coincides at each x with the probability density function obtained from a Gaussian kernel density estimation of $\mathrm{CDS}(i,n,x)$ considered as a function of x: that is, we linearly interpolate the quantity

$$\frac{1}{h*m}\sum_{j=1}^{m} k\left(\frac{x-\mathrm{CDS}(i,n,x_j)}{h}\right), \tag{3.1}$$

where $m=\lfloor\frac{|\mathrm{Chr}(i)|}{4^n+n-2}\rfloor$, $k(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$, and the bandwidth h is selected according to Silverman's rule [39].

The selection of the window size in $P^{\mathrm{hs}}(i,n,x,\varphi)$ exhibits the typical trade-off between sensitivity and specificity: a smaller window size allows for a finer approximation of the CDS distribution, but exhibits a higher sensitivity to fluctuations in nucleotide composition. We focus on a window size of $65,543$ ($n=8$), as this seems to achieve a good balance. This corresponds to dividing $\mathrm{Chr}(1)$ into roughly 3700 non-overlapping windows.

After fixing i and n, we utilize the Nelder-Mead [29] method to maximize the correlation between $P^{\mathrm{hs}}(i,n,x,\varphi)$ and $\mathrm{CDS}(i,n,x)$ with respect to potentials φ which depend on 3 symbols and whose parameters sum to one. The precision threshold for the convergence of this heuristic maximization technique was set to 10^{-6} and convergence was typically achieved in 4000 steps of the algorithm. We denote the potential thus obtained on the i^{th} chromosome by $\varphi^i_{\max}$.

3.5. **Results.** For each chromosome, we obtain a potential $\varphi^i_{\max}$ for which $CDS(i,8,x)$ and $P^{\mathrm{hs}}(i,8,x,\varphi^i_{\max})$ display very strong positive correlation. The value of the Pearson correlation coefficient on each chromosome is shown in figure 1, and is above 0.9 in all cases. Figure 1 also demonstrates that topological entropy is not a good estimator of coding sequence density. This is unsurprising since we have no theoretical reason to expect correlation between entropy and coding sequence density since multiple intron and exon regions may be contained in windows of this size. The parameter values for each $\varphi^i_{\max}$ can be found at

`http://www.math.psu.edu/koslicki/potentials.xls`

We also provide in figure 2 a plot of the standardized values of both $\mathrm{CDS}(5,8,x)$ and $P^{\mathrm{hs}}(5,8,x,\varphi^5_{\max})$ to show the goodness of fit, and overlay these plots on the Ensemble Genome Browser [22] histogram of known genes.

3.6. **Comparison of the potentials $\varphi^i_{\max}$.** Let $\mathbf{r}^i = (r^i_1, r^i_2, \ldots, r^i_{64})$ represent the 64 parameters of the potential $\varphi^i_{\max}$. We show that the parameters for $\varphi^i_{\max}$ exhibit a consistent codon bias by demonstrating that the probability vectors $\mathbf{r}^i$ are relatively close in the standard Euclidean metric. Figure 3 is a plot of the pairwise Euclidean distances between each of the chromosomes. We have

$$\max_{i,j\in\{1,\ldots,24\}} d(\mathbf{r}^i,\mathbf{r}^j) = .319 \qquad \text{and} \qquad \operatorname*{mean}_{i,j\in\{1,\ldots,24\}} d(\mathbf{r}^i,\mathbf{r}^j) = .203$$

The sex chromosomes X and Y are clear outliers, so focusing on the autosomes, these values improve to

$$\max_{i,j\in\{1,\ldots,22\}} d(\mathbf{r}^i,\mathbf{r}^j) = .284 \qquad \text{and} \qquad \operatorname*{mean}_{i,j\in\{1,\ldots,22\}} d(\mathbf{r}^i,\mathbf{r}^j) = .195$$

This is relatively close against a maximum possible distance of $\sqrt{2}$.

In [24], it was observed that the sex chromosomes exhibit a distinctly different entropy distribution than the autosomes. This observation coincides with the fact that the sex chromosome potentials were furthest from the autosomal potentials. Interestingly, the potential $\varphi^7_{\max}$ corresponding to chromosome 7 was similarly distant from the other chromosomes. This is consistent with the fact that chromosome 7 contains many regions identical to the sex chromosomes (of $30,000$ non-overlapping sequences of length $5,000$ from Chr(7), over 77% matched identically with a sequence in chromosome Y).

3.7. **The best choice of potential for CDS density estimation.** We make a 'canonical' choice of potential for estimation of CDS density on the human genome by taking a suitable average of the potentials $\varphi^i_{\max}$. There are various natural ways to do this, each yielding qualitatively similar results. The resulting potential is meaningful because, as shown in §3.6, the individual potentials are close to each other.

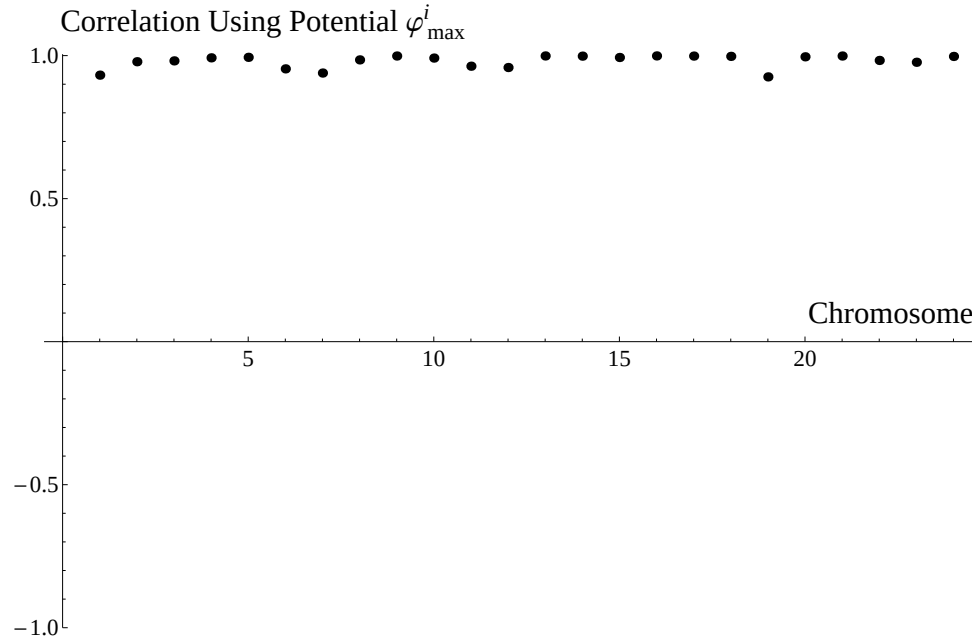
Definition 3.3. *We define a 'canonical' potential for detecting coding sequence density in the human genome which we denote by φ_{hs}. For each codon w, we let*

$$\varphi^0_{\mathrm{hs}}(w) := \mathrm{median}\{\varphi^1_{\max}(w), \ldots, \varphi^{24}_{\max}(w)\},$$

and define

$$\varphi_{\mathrm{hs}} := \frac{\varphi^0_{\mathrm{hs}}}{\|\varphi^0_{\mathrm{hs}}\|}$$

FIGURE 1. Correlation between pressure and CDS density



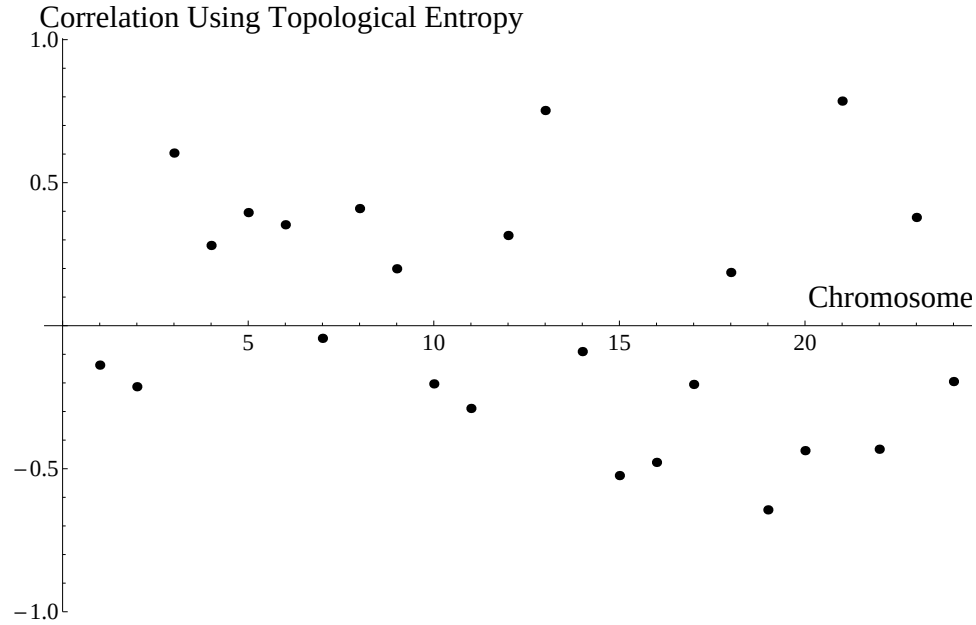
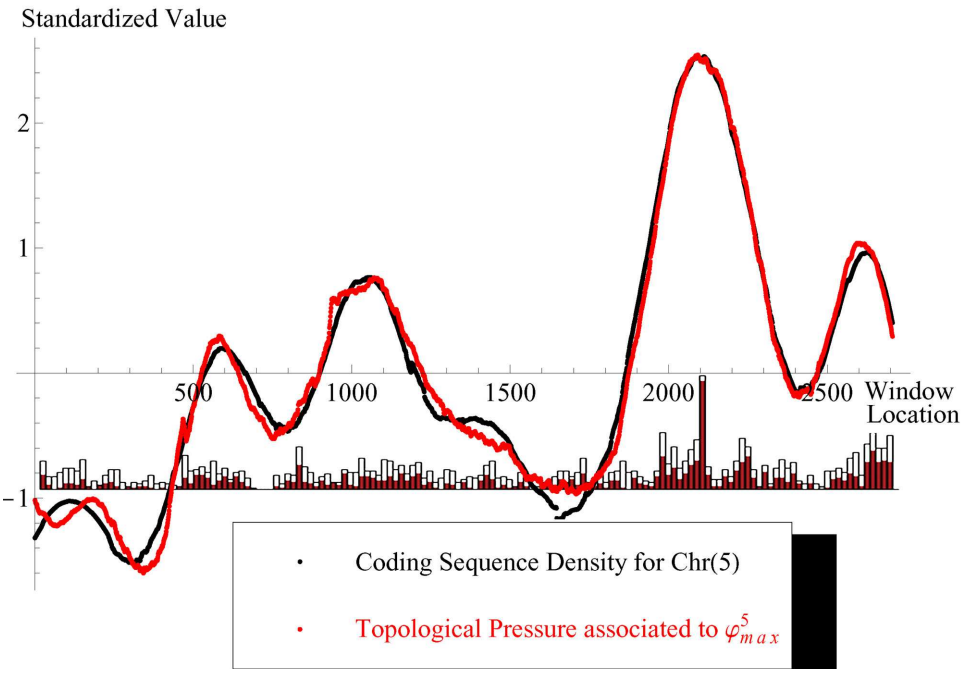


FIGURE 2. Coding sequence density, pressure, and Ensembl known CDS histogram



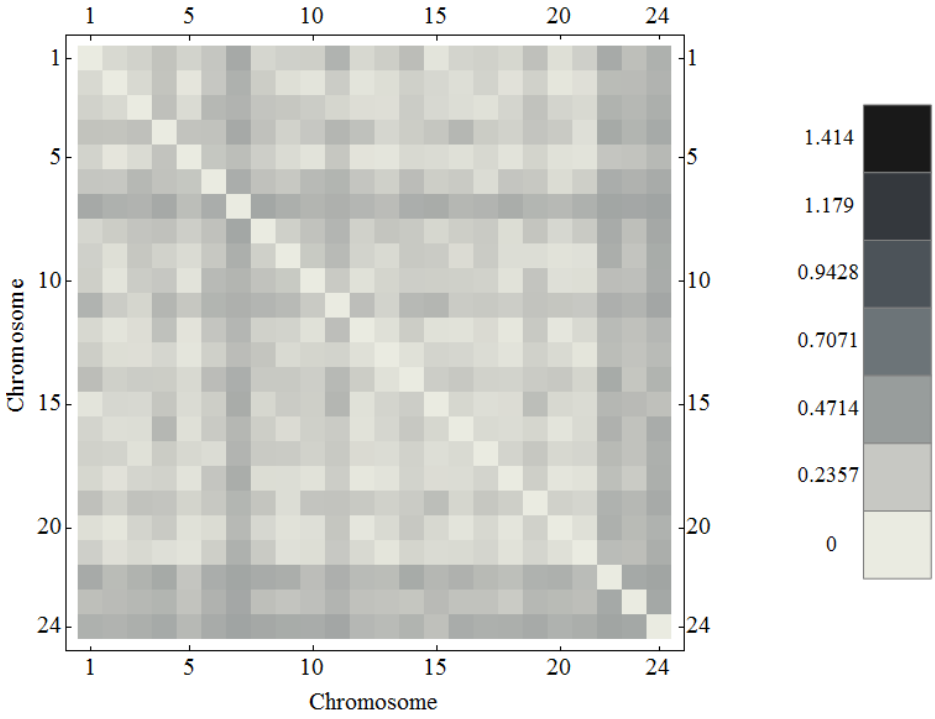
Other natural ways to obtain the 'canonical' potential would be to take the mean of the parameter values of each $\varphi^i_{\max}$, or to take the median/mean after omitting the outlying chromosomes X, Y and 7 from the data set. Each of these approaches yields a very similar potential. Alternatively, we can perform the maximization procedure on the sequence formed by concatenating all the autosomes. This approach yields a potential which is close to φ_{hs} (Euclidean distance less than .148) and qualitatively identical.

We include a visualization of the parameters of φ_{hs} in figure 4. The results of the correlation between the topological pressure $P(i, 8, x, \varphi_{\text{hs}})$ and $\text{CDS}(i, 8, x)$ for the autosomes are contained in figure 5.

3.8. **Analysis of parameter values for φ_{hs}.** We give an in-depth analysis of our canonical potential. As can be observed from figure 4, it is clear that φ_{hs} exhibits a distinct codon bias. We summarize and attempt to explain some of the most distinctive features of φ_{hs}:

- The codons UCG, CCG, UAC, CGC, CGG, AGG and GGC are the most heavily weighted, and the codons CGU, ACU, GCG, UAU, UGA have quite strong weightings.
- Codons which contain the pair CG or GC tend to be highly weighted (for example UCG, CCG, GCG, CGC, CGG, GCC). This is explained by the well known connection between GC content and CDS density. However, we will see in §3.10 that basing a potential on

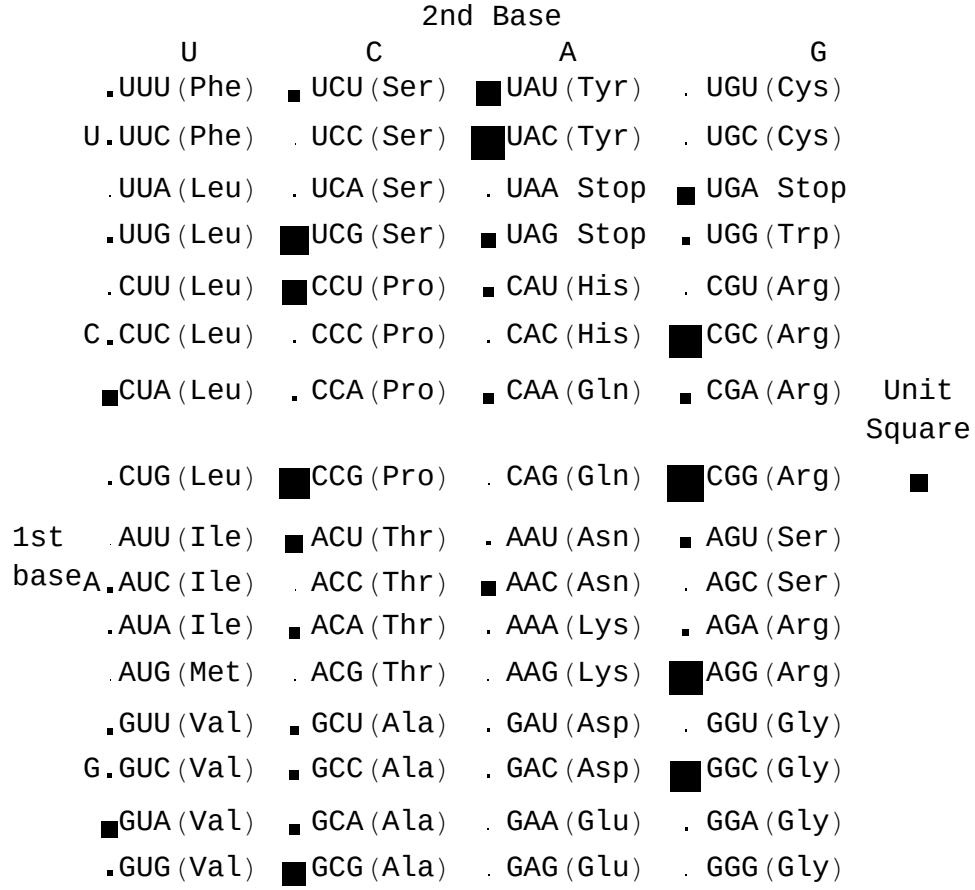
FIGURE 3. Pairwise Euclidean distance of the parameter values for $\varphi^i_{\max}$. The darker the square in position (i, j), the greater the distance between $\mathbf{r}^i$ and $\mathbf{r}^j$.



GC content alone is not sufficient for accurate estimation of CDS density.

- According to the weights in φ_{hs}, the expected GC content of a sequence is 58.4%, which is moderately high since it was shown in [35] that average GC content for a 100-kb segment of the human genome is between 35% and 60%. We calculate expected GC content by the formula
$$\sum_{w \in \mathcal{A}^3} \varphi_{\text{hs}}(w)(\mathrm{N}_G(w) + \mathrm{N}_C(w))/3,$$
where $\mathrm{N}_G(w)$ denotes the number of times the letter G appears in the word w (similarly for $\mathrm{N}_C(w)$). This supports the commonly held notion that high GC content corresponds to high coding sequence density in the human genome [3, 18].
- The start codon (AUG) is weighted near zero. This may indicate that from a large scale perspective, start codons provide too weak a signal to utilize in estimating CDS density.
- The stop codons UGA, UAG, UAA exhibit a decreasing order of significance. This reflects the observation contained in [41] that UGA

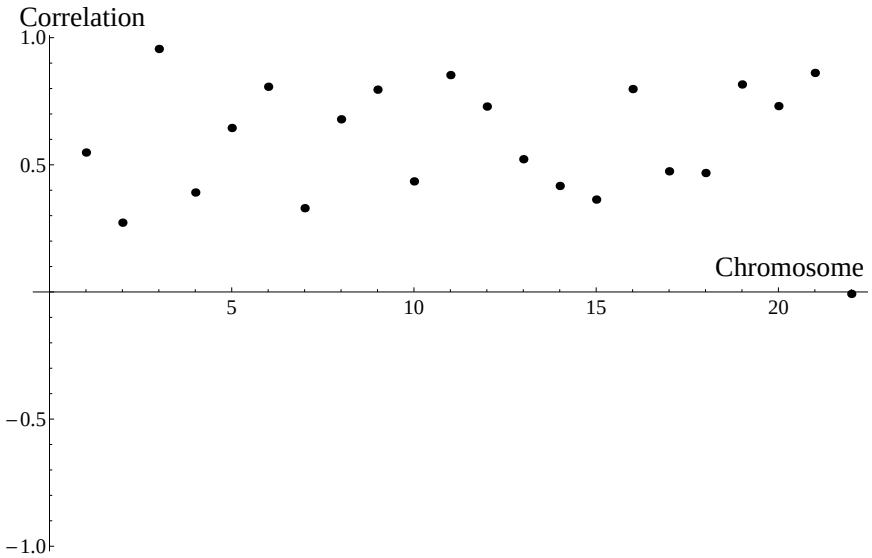
FIGURE 4. Plot of 50 times the parameter values of φ_{hs}. The area of a square is equal to the corresponding potential value.



is utilized most frequently to terminate transcription in the human genome. Furthermore, this pattern of decreasing importance of stop codons reflects the alternate decoding of stop codons (see [11], [40], [45], etc.). In particular, the two stop codons that can be alternately transcribed (UGA and UAG into Selenocysteine and Pyrrolysine respectively) are weighted much more strongly than UAA.

- Codons made up of a single repeating nucleotide receive consistently low weights. This can be explained by the presence of long repetitive regions in non-coding regions.
- We analyzed a number of physical properties associated to amino acids and codons (e.g. acidity, polarity, hydropathy, etc.), and found a weak (.293) but statistically significant ($p < .025$) correlation between the values of φ_{hs} and heat of combustion of the corresponding codons. We are not aware of any results in the literature which would give a theoretical basis for this observation.

FIGURE 5. Plot of Pearson correlation coefficient between the coding sequence density of the human genome and the topological pressure associated to the potential φ_{hs}.



- Synonymous codons may receive very different weightings. For example, among the codons which specify Glycine, GGC is strongly weighted but GGU, GGA and GGG are all weighted near zero.
- Of the five amino acids with four-fold degenerate sites, distinct codon bias is observed: each amino acid with a four-fold degenerate site shows a clear bias towards a nucleotide in the third site (with the exception of Proline where two particular codons are favored). This corresponds with the observations of previous comparative studies, for example [5, 34], where it was observed that there exists selectively driven codon usage at four-fold degenerate sites for mammals (with a weak bias towards C). Our study suggests that any of the four nucleotides may be favored in the third position (A for Val, G for Ala, C for Gly, U for Thr).
- Amino acids with twofold degenerate sites seem typically to carry similar weightings. For example, GAA and GAG both have negligible weightings, while UAU and UAC are both weighted quite strongly. The mean variance of the weighting at twofold degenerate sites was 4.7×10^{-5} while the mean variance over all amino acids was 1.6×10^{-4}.
- For most amino acids, either exactly one codon is weighted strongly (Leu, Val, Ser, Thr, Ala, His, Gly) (or at least more strongly than the others (His, Gln, Asn)), or no codons are weighted strongly (Phe, Ile, Met, Lys, Asp, Glu, Cys, Trp). A notable exception is

Arginine where three out of its six synonymous codons are weighted strongly. This may suggest that Arginine has a particularly important role. This could reflect the evolutionary pressure exerted on Arginine as observed in [23], where it is noted that Arginine has a much lower frequency of appearance than expected. Recently, in [25] it was shown that in yeast, preferential synonymous codon usage for Arginine greatly affects expression levels via influencing translational efficiency. Our results may indicate that a similar phenomenon occurs in the human genome, as this would be another explanation for the strong weighting of Arginine.

3.9. **Selecting potentials using intron/exon density.** Many single sequence techniques for measuring the coding potential of DNA sequences are based upon frequencies of codons or n-mers in known intronic and exonic regions [1, 9, 10, 21]. We can use this principle to write down potentials $\varphi^i_{\text{intron}}$ and φ^i_{exon} which are based simply on the frequency of codons in the intron (or exon) sequences.

More precisely, we let Introns(i) denote the collection of all segments of chromosome i which correspond to known intron regions. For each $w \in$ Introns(i), we let $\mathrm{N}_v(w)$ denote the number of times a given codon v appears in w, and we note that the total number of codons (with overlap) in w is $|w| - 2$. We define a potential $\varphi^i_{\text{intron}}$ by assigning each codon a weight by the formula

$$\varphi^i_{\text{intron}}(v) := \sum_{w \in \text{Introns}(i)} \frac{\mathrm{N}_v(w)}{|w| - 2}. \tag{3.2}$$

We define potentials φ^i_{exon} analogously, using the frequencies of codons that appear in known exon regions of chromosome i. Finally, as in definition 3.3, let $\varphi^0_{\text{exon}}(w) := \text{median}\{\varphi^1_{\text{exon}}(w), \ldots, \varphi^{24}_{\text{exon}}(w)\}$ and define $\varphi_{\text{exon}} := \frac{\varphi^0_{\text{exon}}}{\|\varphi^0_{\text{exon}}\|}$.

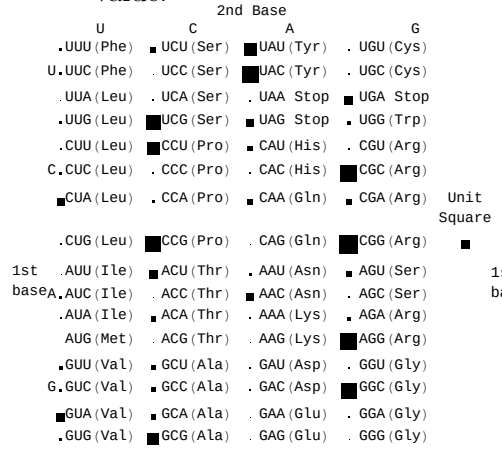
As one would expect, the pressure taken with respect to the potentials $\varphi^i_{\text{intron}}$ (resp. φ^i_{exon}) tends to have significant negative (resp. positive) correlation with the coding sequence density: see figure 7. The mean correlation between $P(i, 8, x, \varphi^i_{\text{intron}})$ and $\text{CDS}(i, 8, x)$ was $-.531$. The mean correlation between $P(i, 8, x, \varphi^i_{\text{exon}})$ and $\text{CDS}(i, 8, x)$ was .376. While this clearly shows a correlation, it is significantly weaker than that obtained using the potentials $\varphi^i_{\max}$. We conclude that potentials which are based simply on frequencies of occurrence of codons in intron/exon regions are useful to an extent, but that more sophisticated potentials, such as φ_{hs}, yield much better results. See figure 6 for a comparison of the potentials φ_{hs} and φ_{exon}.

3.10. **Selecting potentials to detect GC content.** We investigate the pressure of the family of potentials introduced in §2.4:

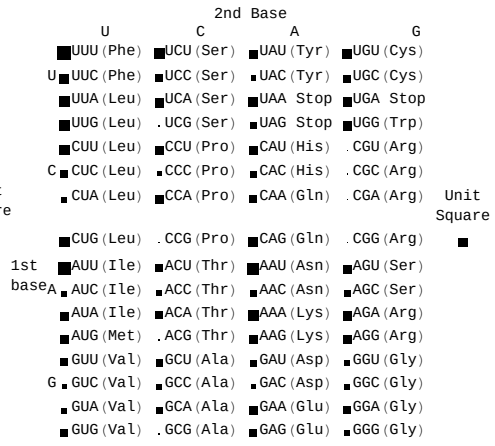
$$\varphi_t = \mathbf{1}_A + \mathbf{1}_T + t(\mathbf{1}_G + \mathbf{1}_C),$$

where $t > 0$. It is a commonly held notion that high GC content corresponds to high coding sequence density in the human genome (see §3.8). We give

FIGURE 6. Visualization of parameter values of potentials. The area of a square is equal to the corresponding potential value.



50×Parameters of φ^{hs}
Overlayed on Genetic Code



50×Parameters of φ_{exon}
Overlayed on Genetic Code

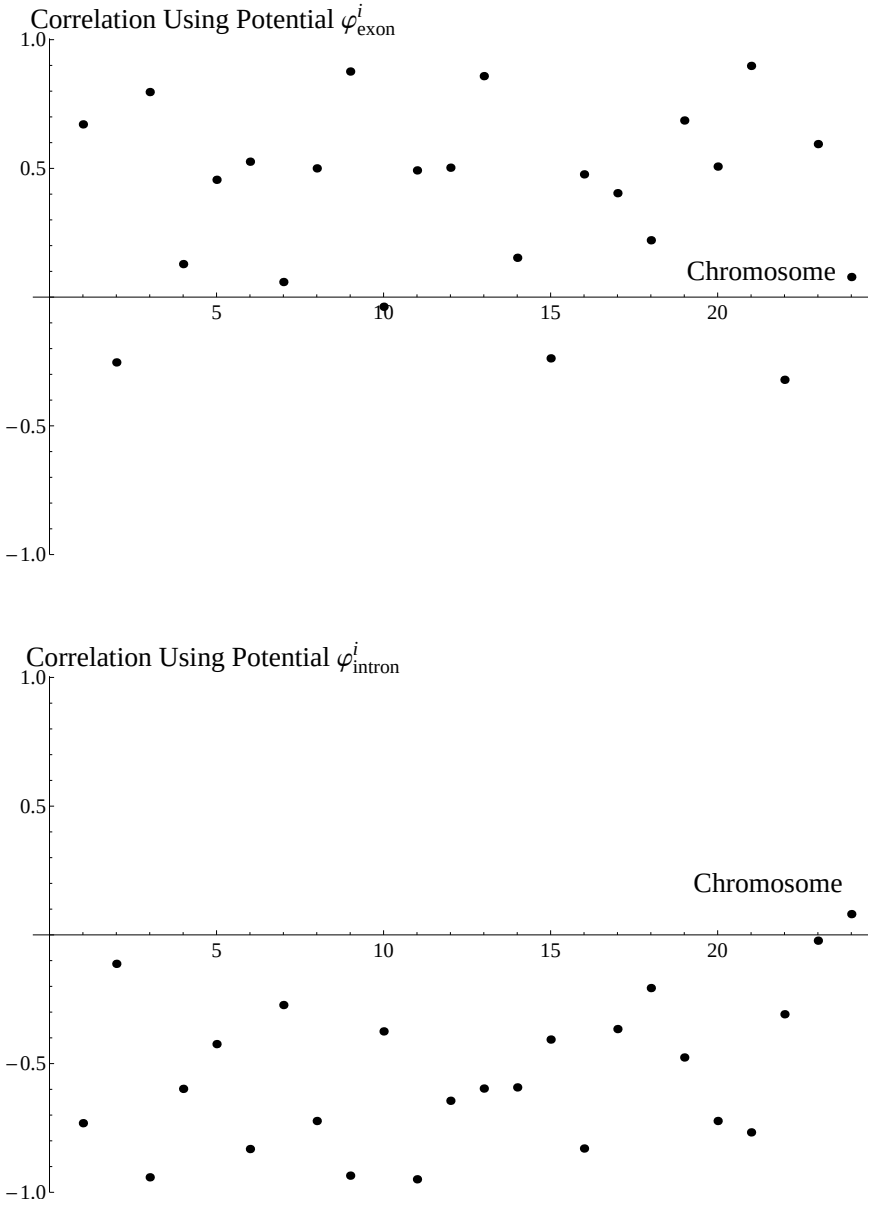
evidence that the link between GC content and CDS density is significant but weak.

For chromosome 1, we find that as t varies, the largest correlation between $P^{\mathrm{hs}}(1, 8, x, \varphi_t)$ and $\mathrm{CDS}(1, 8, x)$ is .138. This maximum is attained (uniquely) when $t = 10.308$. Over all the chromosomes, the maximum correlation of $P^{\mathrm{hs}}(i, 8, x, \varphi_t)$ and $\mathrm{CDS}(i, 8, x)$ has a statistically significant ($p < .0005$) mean of 0.121 with a variance of .00359. This maximum is achieved for a mean parameter value of $t = 10.306$ with a variance of 15.780. The outliers were chromosome 18, which achieves maximal correlation at $t = 21.246$, and chromosome 15, which achieves maximal correlation at $t = 0$. Excluding these two chromosomes gives essentially the same mean ($t = 10.273$), but a much improved variance of 4.98. These results indicate that potentials based on GC content give a weak positive correlation with CDS density. However, the much higher correlation obtained when using the potential φ_{hs} indicates that considering GC content alone is far from optimal in CDS density estimation.

4. Application of the Potential φ_{hs} to the Mouse Genome

We further illustrate the biological significance of the potential φ_{hs} by examining the correlation between the coding sequence density of the mouse genome and the topological pressure associated to the potential φ_{hs}. Following the setup of section 3.4, we retrieve the mouse genome (build mm9, NCBI build 37) from the UCSC database [15] via Galaxy [17], extract the RefSeq genes, and then define $\mathrm{CDS}^{\mathrm{mm}}(i, n, x)$ and $P^{\mathrm{mm}}(i, n, x)$ for the

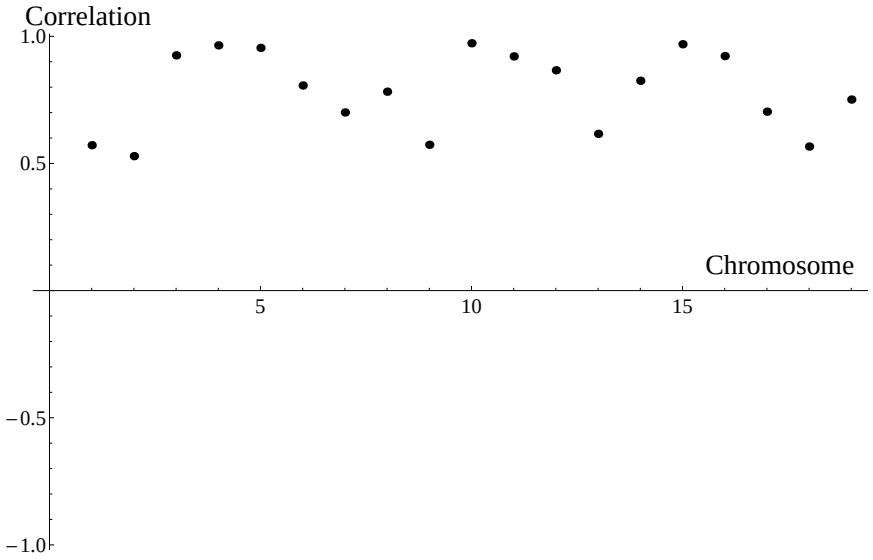
FIGURE 7. Correlation between pressure and coding sequence distribution



mouse autosomes. The results of the correlation between the topological pressure $P^{\text{mm}}(i, 8, x, \varphi_{\text{hs}})$ associated to the potential obtained from the human genome and the coding sequence density of the mouse genome $\text{CDS}^{\text{mm}}(i, 8, x)$ are contained in figure 8. We see a strong positive correlation. This indicates that codon usage in the mouse is similar to codon

usage in humans, and gives further evidence that the potential φ_{hs} genuinely encodes biological information relevant to detecting coding sequence distributions, even across different species.

FIGURE 8. Plot of Pearson correlation coefficient between the coding sequence density of the *Mus Musculus* genome and the topological pressure associated to the potential φ_{hs}.



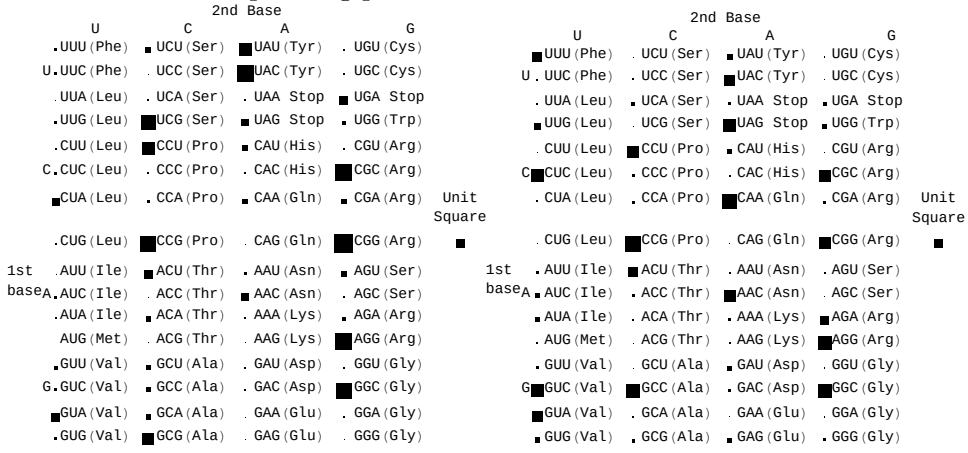
We follow the maximization procedure outlined in §3.4 to obtain potentials $\varphi^i_{\mathrm{mm,max}}$ that maximize the correlation between $P^{\mathrm{mm}}(i, 8, x, \varphi)$ and $\mathrm{CDS}^{\mathrm{mm}}(i, 8, x)$ with respect to φ. Following section §3.7, we average over the potentials $\varphi^i_{\mathrm{mm,max}}$ to obtain a 'canonical' potential for the mouse, which we denote by φ_{mm}. In figure 9, we include a visualization of the parameter values for φ_{mm} and φ_{hs} to demonstrate the similarities between them. It will be an interesting project to carry out this procedure for a much larger collection of species and investigate the similarities and differences between the potentials that are selected for each species.

5. Equilibrium measures and DNA

As mentioned in the introduction, an important area of research is to develop single sequence measures that effectively distinguish between short coding sequences and short non-coding sequences. Here, we utilize ergodic theory to develop such a measure.

Given a locally constant function ψ on $\mathcal{A}^{\mathbb{N}}$, the theory of thermodynamic formalism gives us a means of selecting a Markov measure μ_ψ, known as the equilibrium measure for ψ. We adapt this theory to the case of finite sequences. The measure thus obtained reflects the properties of the function

FIGURE 9. Plot of parameter values of median potentials for human and mouse respectively. The area of a square is equal to the corresponding potential value.



50× Parameters of φ_{hs} Overlayed on Genetic Code

50× Parameters of $\varphi_{\text{mm,max}}$ Overlayed on Genetic Code

ψ. We carry out this procedure for our canonical potential φ_{hs} and obtain a measure that is effective for the analysis of relatively short segments of DNA sequences. We give a brief review of the theory of equilibrium measures in the case of potentials which depend on 3 symbols, and show that the measure thus obtained is meaningful in the finite setting also. We then give numerical results to demonstrate that our measure can distinguish between coding and non-coding sequences with a reasonably high probability of success.

5.1. **Constructing Markov measures from potentials.** We are primarily concerned with functions that depend on 3 symbols, using the alphabet $\mathcal{A} = \{A, C, T, G\}$. That is, we consider potentials $\psi = \log \varphi$, where

$$\varphi = \sum_{w \in \mathcal{A}^3} t_w \mathbf{1}_w, \tag{5.1}$$

so that each $t_w > 0$ is a parameter associated to the word $w \in \mathcal{A}^3$. We review how this this function defines a Markov measure with memory 2 on Σ. The presentation here is a special case of more general expositions given in [2, 4, 28, 31, 32, 42]. Let $\mathcal{B} = \{A, C, G, T\}^2$. Enumerate $\mathcal{B}$ by some natural ordering. For example, let

$$w_1 = AA, w_2 = AC, w_3 = AG, w_4 = AT, w_5 = CA, \ldots, w_{16} = TT$$

Define a $1-0$ square matrix S of dimension 16 as follows. Let $S_{ij} = 1$ if and only if the second letter in w_i is the same as the first letter in w_j. Otherwise, set $S_{ij} = 0$.

We now use the potential ψ to define a non-negative matrix M of dimension 16 as follows. Let $g_{ij} = \log t_w$, where if $w_i = IJ$, and $w_j = JK$, then $w = IJK$. Let $g(i,j) = 0$ if the second letter in w_i is not the same as the first letter in w_j. We define M by

$$M_{ij} = S_{ij} e^{g(i,j)}. \tag{5.2}$$

The Perron-Frobenius theorem gives a maximal eigenvalue $\lambda > 0$ and a strictly positive vector r such that

$$Mr = \lambda r.$$

Now define a matrix P of dimension 16 by

$$P_{ij} = \frac{M_{ij} r_j}{\lambda r_i}. \tag{5.3}$$

It is easy to check that P_{ij} is a stochastic matrix and that there is a unique probability vector p so that $pP = p$. More explicitly, p_i is given by normalizing the vector $l_i r_i$, where l is a strictly positive left eigenvector for M. For $a, b, c \in \mathcal{A}$, let $p(ab) = p_i$ when $ab = w_i$, and let $P(ab, bc) = P_{ij}$ when $ab = w_i$ and $bc = w_j$. We use the pair (p, P) to define a measure as follows.

Definition 5.1. *We define a probability measure μ_ψ on $\mathcal{A}^{\mathbb{N}}$, or $\mathcal{A}^n$ for any fixed $n \geq 3$, by the formula*

$$\mu_\psi([x_1 \dots x_k]) = p(x_1x_2)P(x_1x_2, x_2x_3)P(x_2x_3, x_3x_4) \dots P(x_{k-2}x_{k-1}, x_{k-1}x_k).$$

We call the measure μ_ψ the equilibrium measure for ψ.

5.2. **Properties of the equilibrium measure.** We recall the classical theory from dynamical systems which explains the importance of μ_ψ. First, we recall the definition of topological pressure for the full shift.

Definition 5.2. *The topological pressure of ψ on the full shift Σ over an alphabet $\mathcal{A}$ is defined to be:*

$$P(\Sigma, \psi) = \lim_{n\to\infty} \frac{1}{n} \log \left(\sum_{u \in \mathcal{A}^n} \exp \sum_{i=0}^{n-1} \psi(\sigma^i u) \right).$$

The following result gives the fundamental relationship between pressure and invariant measures [32, 42].

Theorem 5.1 (Variational Principle)**.** *$P(\Sigma, \psi) = \sup_m \left\{ h_m + \int \psi dm \right\}$, where the supremum is taken over all σ-invariant probability measures on Σ, and h_m denotes the measure theoretic entropy, given by*

$$h_m = \lim_{n\to\infty} -\frac{1}{n} \sum_{w \in \mathcal{A}_n} m([w]) \log m([w]).$$

The variational principle illustrates the trade off between structure and complexity which is detected by pressure. Pressure effectively balances the inherent randomness in a sequence while still reflecting the emphasis encoded by the potential φ. Pressure simultaneously maximizes entropy (which is

maximized by the uniform measure) and the average value of the potential (the integral itself is maximized by a Dirac measure). A measure achieving the supremum in the variational principle is called an *equilibrium measure* for ψ. The following result, proved in [32, §4], tells us that the measure constructed in the previous section is indeed an equilibrium measure.

Theorem 5.2. *The Markov measure $\mu = \mu_\psi$ is the unique equilibrium measure for ψ and*

$$P(\Sigma, \psi) = h_\mu + \int \psi d\mu = \log \lambda,$$

where λ is the Perron-Frobenious eigenvalue of the matrix (5.2).

The relationship between ψ and μ_ψ is captured by the *Gibbs property*, established in [4, 31].

Theorem 5.3 (Gibbs property)**.** *For $\psi = \log \varphi$ defined as in* (5.1) *and any $w \in \mathcal{A}^n$,*

$$\mu_\psi([w]) \asymp \exp\{-nP(\Sigma, \psi) + \sum_{i=1}^{n-2} \psi(w_i^{i+2})\},$$

where $a_n \asymp b_n$ means there exists a constant $C > 1$ so that $C^{-1} \leq a_n/b_n \leq C$ for all n.

Thus, if we normalise ψ so that $P(\Sigma, \psi) = 0$ (which is done by taking a suitable multiple of φ), then

$$\mu_\psi([w]) \asymp \prod_{i=1}^{n-2} \varphi(w_i^{i+2}). \tag{5.4}$$

5.3. **Relationship between the equilibrium measure and pressure for finite sequences.** We apply the theory developed in §5.2 to finite sequences. The proof of the following result is similar to that of [42, Theorem 7.30].

Theorem 5.4. *When ψ depends on 3 symbols, $P_{max}(\psi, n) = \log \|M^{n-2}\|^{1/n}$, where M is the matrix constructed in* (5.2)*. Since $\|M^{n-2}\|^{1/n}$ converges to λ exponentially fast as $n \to \infty$, then for large n, $P_{max}(\psi, n)$ is very close to* $\log \lambda$.

Thus, the number $\log \lambda$ is still important for finite sequences. The formula (5.4) reveals the meaning of the measure μ_ψ. Sequences which have a relatively high frequency of words $w \in \mathcal{A}^3$ where t_w is large, and a relatively small frequency of words $w \in \mathcal{A}^3$ where t_w is small, will have relatively large measure. This gives a theoretical underpinning for using μ_ψ to predict coding sequence density.

5.4. **An equilibrium measure for CDS density estimation.** We show that the equilibrium measure has practical applications to distinguishing between coding and non-coding DNA sequences. Recall that in §3 we found a potential φ_{hs} for which the pressure of human DNA segments of length $65,536$bp has strong positive correlation with the coding sequence density. We now show that the equilibrium measure associated to $\log \varphi_{\text{hs}}$, which we denote by μ_{hs}, can distinguish between coding and non-coding DNA sequences with a reasonable degree of success.

The advantage of using the measure μ_{hs} rather than pressure associated to φ_{hs} is that the measure is effective in analyzing relatively short DNA sequences (10bp-5000bp). Indeed, when ψ depends on 3 symbols, pressure is only defined for sequences of length at least $4^4 - 4 = 251$, and only becomes an effective tool for much longer sequences where the noise inherent in the calculation of pressure is effectively suppressed. While the equilibrium measure is a cruder tool than the pressure, it is nevertheless effective for analyzing shorter sequences where the pressure is unavailable.

To demonstrate this phenomena, we show that the measure μ_{hs} can partially distinguish between a randomly selected assortment of intron and exon sequences that are more than an order of magnitude shorter than the sequences on which pressure was evaluated. We randomly select $5,000$ intron sequences and $5,000$ exon sequences from Chr(1), each of length $5,000$bp. These sequences are completely un-preprocessed: no information such as ORF's, stop/start codons or repeat masking is utilized. As expected, the measure μ_{hs} reflects the properties of the potential φ_{hs}: exon sequences are typically weighted more heavily than intron sequences. This is demonstrated by figure 10, which shows the histogram of $\log(\mu_{\text{hs}})$ evaluated on the test sequences. We also include the ROC curve (true positive rate vs. false positive rate) associated to μ_{hs}. The area under the curve (AUC) is 0.722.

FIGURE 10. Histogram of $\log(\mu_{\text{hs}})$ evaluated on the test set of introns and exons

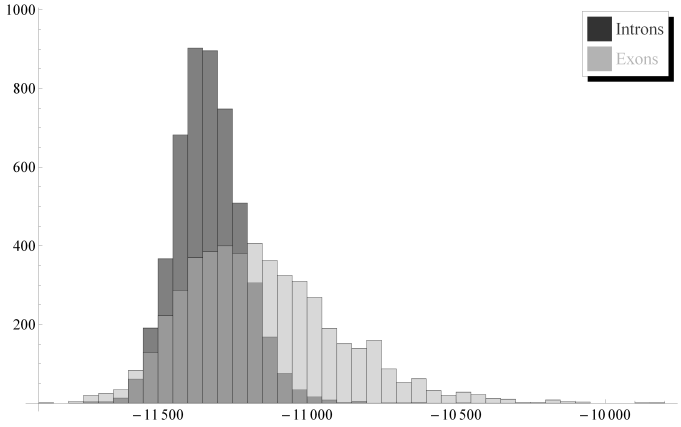
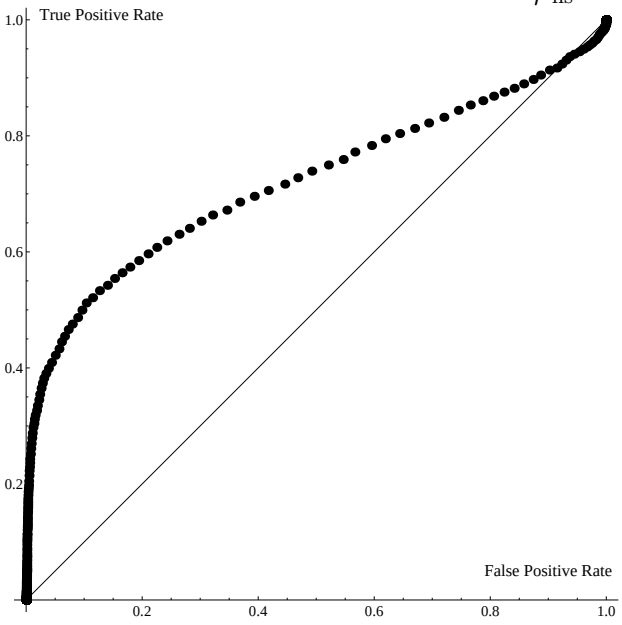


FIGURE 11. ROC curve associated to μ_{hs}.



6. Conclusion

We have introduced a definition of topological pressure for finite sequences inspired by, and related to, the classical definition from ergodic theory. We have applied this definition to DNA sequences in four distinct fashions. First, we obtained a potential that effectively estimates the distribution of coding sequences across the human genome. Second, we gave a detailed analysis of this potential to give new evidence about which codons are most important in coding sequence density estimation. Thus, pressure can be used as a tool for the study of synonymous codon usage. Our analysis effectively measures which codons are most important and not simply most frequently appearing. Third, we used topological pressure to compare coding sequence density in the human and the mouse genome, giving evidence via pressure that codon usage is similar across both species. Lastly, we derived the equilibrium measure associated to a particular potential and showed that this can be used to distinguish between relatively short reads of coding and non-coding sequences.

This study has indicated that topological pressure may help elucidate the nuanced problems of mammalian codon bias. Since topological pressure does not rely on a particular statistical perspective but is motivated by a rigorous implementation of a well developed mathematical theory, we expect that our approach will yield many further applications in genomic analysis in the future. We expect that the inclusion of pressure in comparative studies will contribute to the understanding of the relationship between

codon bias and gene expression levels. Furthermore, the ability of equilibrium measures to succinctly encapsulate information obtained on very large scales indicates the usefulness of pressure for the development of measures of coding potential for short DNA sequences.

References

[1] H. Akashi. Gene expression and molecular evolution. *Curr Opin Genet Dev*, 11(6):660–666, 2001.

[2] V. Baladi. *Positive transfer operators and decay of correlations*, volume 16. World Scientific, 2000.

[3] G. Bernardi. The isochore organization of the human genome. *Annu Rev Genet*, 23:637–661, 1989.

[4] R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, volume 470 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin-New York, 1975.

[5] J.-V. Chamary and L. D. Hurst. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol*, 21(6):1014–23, 2004.

[6] J. V. Chamary and L. D. Hurst. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol*, 6(9):R75, 2005.

[7] J. V. Chamary, J. L. Parmley, and L. D. Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, 7(2):98–108, 2006.

[8] S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA*, 101(10):3480–5, 2004.

[9] J. M. Comeron and M. Aguadé. An evaluation of measures of synonymous codon usage bias. *J Mol Evol*, 47(3):268–74, 1998.

[10] T. M. Creanza, D. S. Horner, A. D'Addabbo, R. Maglietta, F. Mignone, N. Ancona, and G. Pesole. Statistical assessment of discriminative features for protein-coding and non coding cross-species conserved sequence elements. *BMC Bioinformatics*, 10 Suppl 6:S2, 2009.

[11] V. A. Doronina and J. D. Brown. When nonsense makes sense and vice versa: Non-canonical decoding events at stop codons in eukaryotes. *Mol Biol*, 40(4):654–663, 2006.

[12] M. Dos Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, 32(17):5036–44, 2004.

[13] A. Fedorov, S. Saxonov, and W. Gilbert. Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res*, 30(5):1192–7, 2002.

[14] J. W. Fickett and C. S. Tung. Assessment of protein coding measures. *Nucleic Acids Res*, 20(24):6441–50, 1992.

[15] P. Fujita, B. Rhead, A. Zweig, A. Hinrichs, D. Karolchik, M. Cline, M. Goldman, G. Barber, H. Clawson, A. Coelho, M. Diekhans, T. Dreszer, B. Giardine, R. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. Kuhn, K. Learned, C. Li, L. Meyer, A. Pohl, B. Raney, K. Rosenbloom, K. Smith, D. Haussler, and W. Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(suppl 1):D876–D882, 2011.

[16] F. Gao and C.-T. Zhang. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*, 20(5):673–81, Mar. 2004.

[17] J. Goecks, A. Nekrutenko, J. Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 25(11):R86–R99, 2010.

[18] R. Guigó and J. W. Fickett. Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. *J Mol Biol*, 253(1):51–60, 1995.
[19] Y. G. Gursky and R. S. Beabealashvilli. The increase in gene expression induced by introduction of rare codons into the C terminus of the template. *Gene*, 148(1):15–21, 1994.
[20] R. Hershberg and D. A. Petrov. Selection on codon bias. *Annu Rev Genet*, 42(iv):287–99, 2008.
[21] S. Karlin, J. Mrázek, and A. Campbell. Codon usages in different gene classes of the Escherichia coli genome. *Mol Microbiol*, 29(6):1341–1355, 1998.
[22] e. a. Kersey P. Ensemble genomes: Extending ensembl across the taxonomic space. *Nucleic Acids Research*, 38(suppl. 1):D563–D569, 2010.
[23] J. King. Non-Darwinian evolution. *Science*, 164:788–798, 1969.
[24] D. Koslicki. Topological Entropy of DNA Sequences. *Bioinformatics*, 27(8):1061–1067, 2011.
[25] D. P. Letzring, K. M. Dean, and E. J. Grayhack. Control of translation efficiency in yeast by codon-anticodon interactions. *RNA*, 16(12):2516–28, 2010.
[26] M. F. Lin, A. N. Deoras, M. D. Rasmussen, and M. Kellis. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comp Biol*, 4(4):e1000067, 2008.
[27] M. F. Lin, I. Jungreis, and M. Kellis. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011.
[28] D. Lind and B. Marcus. *An introduction to symbolic dynamics and coding*. Cambridge University Press, 1995.
[29] J. Nelder and R. Mead. A simplex method for function minimization. *Comput J*, 7(4):308, 1965.
[30] J. Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol and Evol*, 19(8):1390–1394, 2002.
[31] W. Parry and M. Pollicott. *Zeta functions and the periodic orbit structure of hyperbolic dynamics*. Number 187-188 in Astérisque. Soc. Math. France, 1990.
[32] W. Parry and S. Tuncel. *Classification problems in ergodic theory*, volume 67 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1982. Statistics: Textbooks and Monographs, 41.
[33] J. B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12(1):32–42, 2011.
[34] A. M. Resch, L. Carmel, L. Mariño Ramírez, A. Y. Ogurtsov, S. A. Shabalina, I. B. Rogozin, and E. V. Koonin. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol*, 24(8):1821–31, 2007.
[35] J. Romiguier, V. Ranwez, E. J. P. Douzery, and N. Galtier. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*, 20(8):1001–9, 2010.
[36] A. H. Rosenberg, E. Goldman, J. J. Dunn, F. W. Studier, and G. Zubay. Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. *J Bacteriol*, 175(3):716–22, 1993.
[37] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–17, Oct. 2007.
[38] W. Seffens and D. Digby. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res*, 27(7):1578–84, 1999.
[39] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.
[40] T. Stadtman. Selenocysteine. *Annu Rev Biochem*, 65:83–100, 1996.

[41] J. Sun, M. Chen, J. Xu, and J. Luo. Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. *J Mol Evol*, 61(4):437–44, 2005.

[42] P. Walters. *An Introduction to Ergodic Theory*, volume 79 of *Graduate Texts in Mathematics*. Springer, New York, 1982.

[43] S. Washietl, S. Findeiss, S. A. Müller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, and N. Goldman. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, 17(4):578–94, 2011.

[44] Wolfram Research. *Mathematica*. Wolfram Research, Inc., Champaign Illinois, 8.0 edition, 2010.

[45] F. Zinoni, A. Birkmann, and W. Leinfelder. Cotranslational insertion of selecocysteine into formate dehydrogenase from Escherichia coli directed by a UGA codon. *Proc Natl Acad Sci*, 84:3156–3160, 1987.

Department of Mathematics, Pennsylvania State University, University Park, PA, 16802

E-mail address: koslicki@math.psu.edu

E-mail address: thompson@math.psu.edu